Linking Motif Sequences with Tale Types by Machine Learning

Nir Ofek[1], Sándor Darányi[2], and Lior Rokach[1]

1 **Department of Information Systems Engineering**
Ben-Gurion University of the Negev
Beer-Sheva, Israel
{nirofek,liorrk}@bgu.ac.il

2 **Swedish School of Library and Information Science**
University of Borås
Borås, Sweden
sandor.daranyi@hb.se

Abstract

Abstract units of narrative content called motifs constitute sequences, also known as tale types. However whereas the dependency of tale types on the constituent motifs is clear, the strength of their bond has not been measured this far. Based on the observation that differences between such motif sequences are reminiscent of nucleotide and chromosome mutations in genetics, i.e., constitute "narrative DNA", we used sequence mining methods from bioinformatics to learn more about the nature of tale types as a corpus. 94% of the Aarne-Thompson-Uther catalogue (2249 tale types in 7050 variants) was listed as individual motif strings based on the Thompson Motif Index, and scanned for similar subsequences. Next, using machine learning algorithms, we built and evaluated a classifier which predicts the tale type of a new motif sequence. Our findings indicate that, due to the size of the available samples, the classification model was best able to predict magic tales, novelles and jokes.

1998 ACM Subject Classification G.3 Probability and statistics, H.2.8 Database applications – Data mining, H.3.1 Content analysis and indexing, H.3.2 Information storage – Record classification, I.2.6 Learning – Parameter learning

Keywords and phrases Narrative DNA, tale types, motifs, type-motif correlation, machine learning

Digital Object Identifier 10.4230/OASIcs.CMN.2013.166

1 Introduction

Digital humanities and the emerging field of cultural analytics implement powerful multidisciplinary metaphors and methods to process texts in unprecedented ways. One of the new concepts is the reference to "narrative genomics" or "narrative DNA" – more and more authors point out similarities between sequences of genetic material building up living material, and those of literary units constituting "memetic", i.e., cultural products whose transmission can be traced by means of population genetics [29].

The idea that canonical sequences of content indicators constitute higher order content units has been pervading biology in the 20th century, and then slowly spilled over to other domains, prominently linguistics. Namely, strictly regulated strings of nucleotides constitute genes whereas canonical strings of genes amount to chromosomes. As a parallel, first the notion of indexing languages as sentence-like sequences of classification tags was born [27], then disciplinary sublanguages as content indicator chains were proposed [15], and finally,

Workshop on Computational Models of Narrative 2013.
Editors: Mark A. Finlayson, Bernhard Fisseni, Benedikt Löwe, and Jan Christoph Meister; pp. 166–182

OpenAccess Series in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

treebanks suggested to manifest the linguistic genome [1], albeit at the cost of giving up canonical sequences for a more loose concept of syntax; that is, at this point in development anything can be considered a genome as long as the expression is sequential and well-formed, i.e., grammatical. As a latest development in this respect, recently Jockers claimed to study the 19th century literary genome of English novels, having extended the genetic metaphor to corpus linguistics to come up with new findings for cultural analytics [18].

Our current endeavor below relates to this latter tradition in the building, although following its own line of thought when considering tale types as canonical motif sequences [5, 6]. By motifs we mean abstracted, generic content tags which summarize segments of the plot. For a more detailed discussion of motifs and related considerations, see, e.g., [4].

As a next step, in this more technical approach to apply methodology from bioinformatics to problems of the literary genome, here we introduce machine learning to reveal the probabilistic scaffolding of tale types in terms of motif content. Whereas the idea is simple – if motifs are condensed expressions of multiple sentence content, then tale types "sum up" motif sequences to yield broader topics –, this first attempt still bears all the hallmarks of a dry run and therefore comes with a caveat.

This paper is organized as follows. Section 2 explains background considerations with special emphasis on formulaity and metadata. Section 3 discusses the research problem. Section 4 offers a brief introduction to sequence mining, with Section 5 outlining the methodology used in the experiment. Section 6 reports the results, whereas Section 7 sums up our conclusions with suggestions for future research. The appendix gives insight in the structure of the tale type catalog and the motif index used in the experiment.

2 Background considerations

2.1 Formulaity as a means of storyline preservation

It has been known for almost a hundred years that the oral communication of folklore texts often applies *formulaity* to help the singer remember his text [24, 25, 22]. Filed under different names, structural and formal investigations of tales [30, 26, 17] and myths, indeed mythologies, have proposed the same approach [21, 23]. Less known is the fact that linguistic evidence points in the same direction: as exemplified by a now famous study in immunology, scientific sublanguages, characteristic of subject areas, may use a formulaic arrangement of content elements in a sequential fashion for the presentation of experiments, results, and their discussion [15]. Formulae as storytelling aids abound in oral literature on all levels and in all genres; consult e.g., [22, 16] for various formulae in the genre of oral epics and [28] for the genre of fairy tale.

Several kinds of formulaity exist, ranging from short canonical phrases such as the *epitheton ornans* in Homeric epics, to longer ones used in orally improvised poetry, including canonical sequences of content elements and leading to story grammars [20, 11] or narrative algebra [12, 13, 14]. We will focus on such sequences only.

To recall, according to the oral-formulaic theory developed by Milman Parry [24, 25] and Albert Lord [22], stock phrases could enable poets to improvise verse called orally improvised poetry. In oral composition, the story itself has no definitive text, but consists of innumerable variants, each improvised by the teller in the act of telling the tale from a mental stockpile of verbal formulas, thematic constructs, and narrative incidents. This improvisation is for the most part subconscious so that texts orally composed will differ substantially from day to day and from teller to teller. The key idea of the theory is that poets have a store of formulas (a formula being 'an expression which is regularly used, under the same metrical conditions,

to express a particular essential idea' [22]), and that by linking these in conventionalized ways, they can rapidly compose verse.

Such linking, however, seems to be pertinent to storytelling in prose as well. The following example displays a chain of motifs which characterize a particular tale type about supernatural adversaries:

> 300 *The Dragon-Slayer.* A youth acquires (e.g., by exchange) three wonderful dogs [B421, B312.2]. He comes to a town where people are mourning and learns that once a year a (seven-headed) dragon [B11.2.3.1] demands a virgin as a sacrifice [B11.10, S262]. In the current year, the king's daughter has been chosen to be sacrificed, and the king offers her as a prize to her rescuer [T68.1]. The youth goes to the appointed place. While waiting to fight with the dragon, he falls into a magic sleep [D1975], during which the princess twists a ring (ribbons) into his hair; only one of her falling tears can awaken him [D1978. 2].
>
> Together with his dogs, the youth overcomes the dragon [B11.11, B524.1.1, R111.1.3]. He strikes off the dragon's heads and cuts out the tongues (keeps the teeth) [H105.1]. The youth promises the princess to come back in one year (three years) and goes off.
>
> An impostor (e.g., the coachman) takes the dragon's heads, forces the princess to name him as her rescuer [K1933], and claims her as his reward [K1932]. The princess asks her father to delay the wedding. Just as the princess is about to marry the impostor, the dragon-slayer returns. He sends his dogs to get some food from the king's table and is summoned to the wedding party [H151.2]. There the dragon-slayer proves he was the rescuer by showing the dragon's tongues (teeth) [H83, H105.1]. The impostor is condemned to death, and the dragon-slayer marries the princess [32].

Square brackets refer to forkings in the plot where alternative motifs can result in valid tale variants (Figure 1).

What matters for our argumentation is that as much as a certain sequence of specific Proppian functions amounts to a fairy tale plot [26], it takes a certain linking of consecutive motifs to constitute a specific tale type. Extracting chains of symbolic content from text in the above sense is the formulaic representation of sentences as proposed by Harris *et al.* [15], bridging the gap between scientific sublanguages and so far unidentified agglomerations of sentences amounting to sequentially linked functions, motifs etc.

2.2 Metadata in folktale research

The case we want to test our working hypothesis on, outlined in Section 5, is the Aarne-Thompson-Uther Tale Type Catalog (ATU), a classification and bibliography of international folk tales [32]. In the ATU, tale types are defined as canonical motif sequences such that motif string A constitutes Type X, string B stands for Type Y, etc. Also, it is important to note that types were not conceived in the void, rather they extract the essential characteristic features of a body of tales from all corners of the world, i.e., they are quasi-formal expressions of typical narrative content, mapped from many to one.

ATU is an alphanumerical, basically decimal classification scheme describing tale types in seven major chapters (animal tales, tales of magic, religious tales, realistic tales [novelle], tales of the stupid ogre [giant, devil], anecdotes and jokes, and formula tales), with an extensive Appendix discussing discontinued types, changes in previous type numbers, new types, geographical and ethnic terms, a register of motifs exemplified in tale types, bibliography and abbreviations, additional references and a subject index.

ATU tale type 300: The Dragon-Slayer.
[B421 B312.2] B11.2.3.1 [B11.10 S262] T68.1 D1975 D1978.2 [B11.11 B524.1.1 R111.1.3] H105.1 K1933 K1932 H151.2 [H83 H105.1]
Sequence variants

B421	B11.2.3.1	B11.10	T68.1	D1975	D1978.2	B11.11	H105.1	K1933	K1932	H151.2	H83
B421	B11.2.3.1	S262	T68.1	D1975	D1978.2	B11.11	H105.1	K1933	K1932	H151.2	H83
B421	B11.2.3.1	B11.10	T68.1	D1975	D1978.2	B524.1.1	H105.1	K1933	K1932	H151.2	H83
B421	B11.2.3.1	S262	T68.1	D1975	D1978.2	B524.1.1	H105.1	K1933	K1932	H151.2	H83
B421	B11.2.3.1	B11.10	T68.1	D1975	D1978.2	R111.1.3	H105.1	K1933	K1932	H151.2	H83
B421	B11.2.3.1	S262	T68.1	D1975	D1978.2	R111.1.3	H105.1	K1933	K1932	H151.2	H83
B421	B11.2.3.1	B11.10	T68.1	D1975	D1978.2	B11.11	H105.1	K1933	K1932	H151.2	H105.1
B421	B11.2.3.1	S262	T68.1	D1975	D1978.2	B11.11	H105.1	K1933	K1932	H151.2	H105.1
B421	B11.2.3.1	B11.10	T68.1	D1975	D1978.2	B524.1.1	H105.1	K1933	K1932	H151.2	H105.1
B421	B11.2.3.1	S262	T68.1	D1975	D1978.2	B524.1.1	H105.1	K1933	K1932	H151.2	H105.1
B421	B11.2.3.1	B11.10	T68.1	D1975	D1978.2	R111.1.3	H105.1	K1933	K1932	H151.2	H105.1
B421	B11.2.3.1	S262	T68.1	D1975	D1978.2	R111.1.3	H105.1	K1933	K1932	H151.2	H105.1
B312.2	B11.2.3.1	B11.10	T68.1	D1975	D1978.2	B11.11	H105.1	K1933	K1932	H151.2	H83
B312.2	B11.2.3.1	S262	T68.1	D1975	D1978.2	B11.11	H105.1	K1933	K1932	H151.2	H83
B312.2	B11.2.3.1	B11.10	T68.1	D1975	D1978.2	B524.1.1	H105.1	K1933	K1932	H151.2	H83
B312.2	B11.2.3.1	S262	T68.1	D1975	D1978.2	B524.1.1	H105.1	K1933	K1932	H151.2	H83
B312.2	B11.2.3.1	B11.10	T68.1	D1975	D1978.2	R111.1.3	H105.1	K1933	K1932	H151.2	H83
B312.2	B11.2.3.1	S262	T68.1	D1975	D1978.2	R111.1.3	H105.1	K1933	K1932	H151.2	H83
B312.2	B11.2.3.1	B11.10	T68.1	D1975	D1978.2	B11.11	H105.1	K1933	K1932	H151.2	H105.1
B312.2	B11.2.3.1	S262	T68.1	D1975	D1978.2	B11.11	H105.1	K1933	K1932	H151.2	H105.1
B312.2	B11.2.3.1	B11.10	T68.1	D1975	D1978.2	B524.1.1	H105.1	K1933	K1932	H151.2	H105.1
B312.2	B11.2.3.1	S262	T68.1	D1975	D1978.2	B524.1.1	H105.1	K1933	K1932	H151.2	H105.1
B312.2	B11.2.3.1	B11.10	T68.1	D1975	D1978.2	R111.1.3	H105.1	K1933	K1932	H151.2	H105.1
B312.2	B11.2.3.1	S262	T68.1	D1975	D1978.2	R111.1.3	H105.1	K1933	K1932	H151.2	H105.1

Figure 1 *300 The Dragon Slayer* as a motif chain and its equally valid story variants.

The numbering of the types runs from 1 to 2399. Individual type descriptions uniformly come with a number, a title, an abstract-like plot mostly tagged with motifs, known combinations with other types, technical remarks, and references to the most important literature on the type plus its variants in different cultures. At the same time, as the inclusion of some 250 new types in the Appendix indicates, tale typology is a comprehensive and large-scale field of study, but also unfinished business: not all motifs in the *Motif Index* [30] were used to tag the types, difficulties of the definition of a motif imposed limitations on its usability in ATU, and narrative genre related considerations related to classification in general had to be observed.[1]

To turn to Thompson's *Motif-Index*, it offers worldwide coverage of folk narrative. As Alan Dundes suggested, in spite of its shortcomings, "It must be said at the outset that the six-volume *Motif-Index of Folk-Literature* and the Aarne-Thompson tale type index constitute two of the most valuable tools in the professional folklorist's arsenal of aids for analysis. This is so regardless of any legitimate criticisms of these two remarkable indices, the use of which serves to distinguish scholarly studies of folk narrative from those carried out by a host of amateurs and dilettantes. The identification of folk narratives through motif and/or tale type numbers has become an international *sine qua non* among bona fide folklorists. For this reason, the academic folklore community has reason to remain eternally grateful to Antti Aarne (1867–1925) and Stith Thompson (1885–1976) who twice revised Aarne's original 1910 *Verzeichnis der Märchentypen*—in 1928 and in 1961—and who compiled two editions of the *Motifindex* (1922–1936; 1955–1958)" [9].

In appendices A and B we give an overview of the structure in ATU and the structure of a sample class of motifs.

[1] Hans-Jörg Uther, personal communication.

3 Research problem

Both transmitted genetic content and transmitted text content undergo variation over time. Genetic variation is called mutation and affects, e.g., nucleotides, amino acids, genes etc. Text variation does not have a specific name. Narrative elements that can vary include motifs (i.e., abstracted, generic content tags which summarize segments of the plot). Motif chains are the "backbones" of tale types, clusters of multilingual texts with related content. As motif insertion, deletion, and crossover were demonstrated to exist in tale types [5], types of mutation known from genetics apparently also occur in storytelling.

With the above observations about formulaity in oral tradition in mind, and to use the terminology of Dawkins [7], given the phenomenon of text variation in folklore, the existence of tale motifs and tale types on a global scale is universal evidence for semantic content resisting erosion, i.e., meme loss. This stability of memetic products invites the study of the relationship between two forms of memes, tag content vs. type content as a classification problem, the relationship between the features of a class and their sum total reflected in a set of documents being a major research issue well beyond folklore research. Therefore to ask about the interplay between genres like animal tales, and the content of motifs which build up such tales so that the result ends up in that genre, justifies one's curiosity. Put another way, this time we were interested in the correlation between two respective semantic fields [31], one described by tale types, the other by thematic motif groups and subgroups.

The stability of content sequences is documented in different corners of text research, lately for example by (Danescu-Niculescu-Mizil *et al.* [3]) who studied the memorability of phrases. In this particular field, Darányi and Forró [5] have shown that motifs are not the ultimate level of tale content available for indexing. In a sample of 219 tale types over 1202 motifs (ATU 300-745A, "Tales of magic" segment), their semiautomated analysis found granularity in ATU on two more levels, in the pattern of motif co-occurrences and in collocated motif co-occurrences, both apparently having been stable enough to resist text variation. On the other hand, Karsdorp *et al.* [19] have indicated that tale types in ATU show reasonably unique motif sequences whose subsequences are hardly ever repeated over different types.

With these considerations in mind, next we briefly introduce the data mining methodology we decided to apply to the problem.

4 Sequence mining by machine learning

Sequential pattern mining is a prevalent data mining approach [8]. The input of the learning process is a set of class labeled sequences (here tale types), which are used to train a model to predict the label of any unlabeled sequence. The learning process for classification uses the information of subsequences derived from the original sequences to discriminate class types. That is performed mainly by calculating a discrimination ratio based on statistics.

Sequence data include sequences of DNA, protein, customer purchase history, web surfing history, and more. Ferreira and Azevedo [10] used sequence mining in conjunction with a machine learning algorithm to classify protein sequences.

5 Methods

We considered motifs as entries in an indexing vocabulary and tale types as the document vectors in a corpus indexed by them, the latter being sparse motif strings which at the same time constitute "sentences", i.e., are predicates about type-specific tale content.

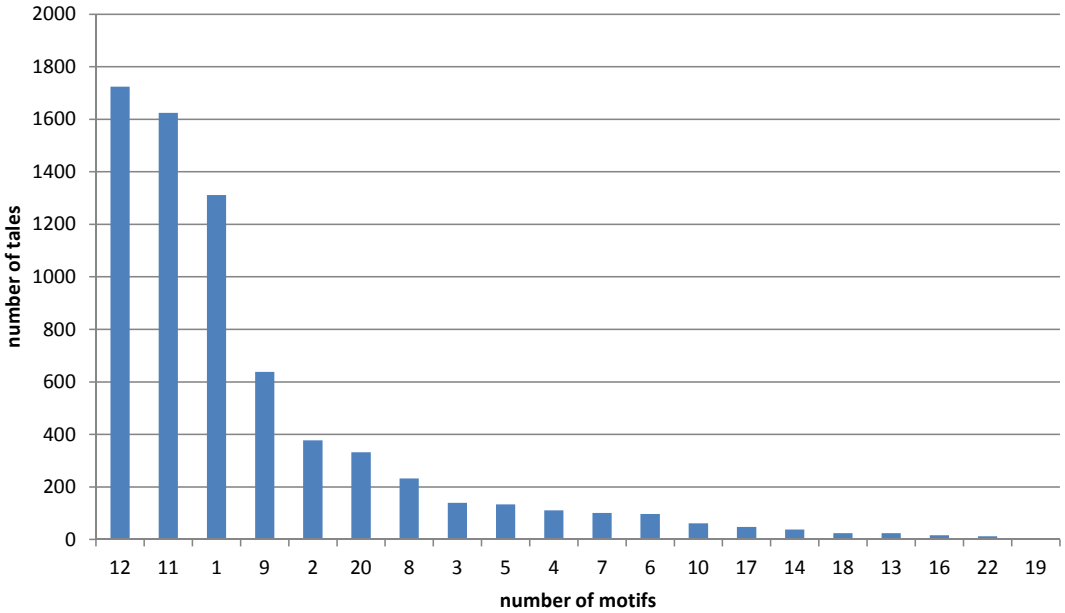

Figure 2 Tale type sequence lengths (in motifs).

We anticipated thematic dependencies between the 7 narrative genres defined as tale types in ATU (i.e., animal tales, tales of magic, religious tales, realistic tales [novelle], tales of the stupid ogre [giant, devil], anecdotes and jokes, and formula tales) vs. the 23 major motif groups in the *Motif Index* (mythological motifs, animals, taboo, magic, the dead, marvels, ogres, tests, the wise and the foolish, deceptions, reversal of fortune, ordaining the future, chance and fate, society, rewards and punishments, captives and fugitives, unnatural cruelty, sex, the nature of life, religion, traits of character, humor, and miscellaneous). The research question was, how do motifs from the above 23 groups constitute sequences resulting in those 7 genres? We assumed that by exploring the dependency structure of motifs vs. tale types, one can unveil the underlying probabilistic underpinnings of storyline construction.

5.1 Material

We used 94% of the complete ATU for this first experiment, i.e., out of the 2399 types we worked with 2249. The remaining 6% were left out from preprocessing because of their non-standard motif notation in the types, e.g., also containing running text in square brackets. Figure 2 shows the frequency of tale type length in terms of number of motifs in the string. Figure 3 displays the number of tale types and subtypes per genre.

It is an open question if due to text erosion or because of still being in a nascent stage, but many of the tale types consist of a single motif only. This undermines the very notion of tale type as a motif sequence [9]. In ATU, one-motif narratives are typical for anecdotes and jokes, formula tales and animal tales, whereas they are least characteristic for tales of magic, with the other genres statistically placed between them. Contrary to Karsdorp *et al.* [19], we feel that tale type is a bicomponential concept, having to satisfy both a formal and a topical constraint, and where the formal aspect, i.e., being a string, is met to a minimum only, topicality still prevails and accounts for the existence of genres grouping short texts; furthermore nothing prevents one from concatenating them in order to generate new narrative types with a higher dose of adventure than in anectodes etc.

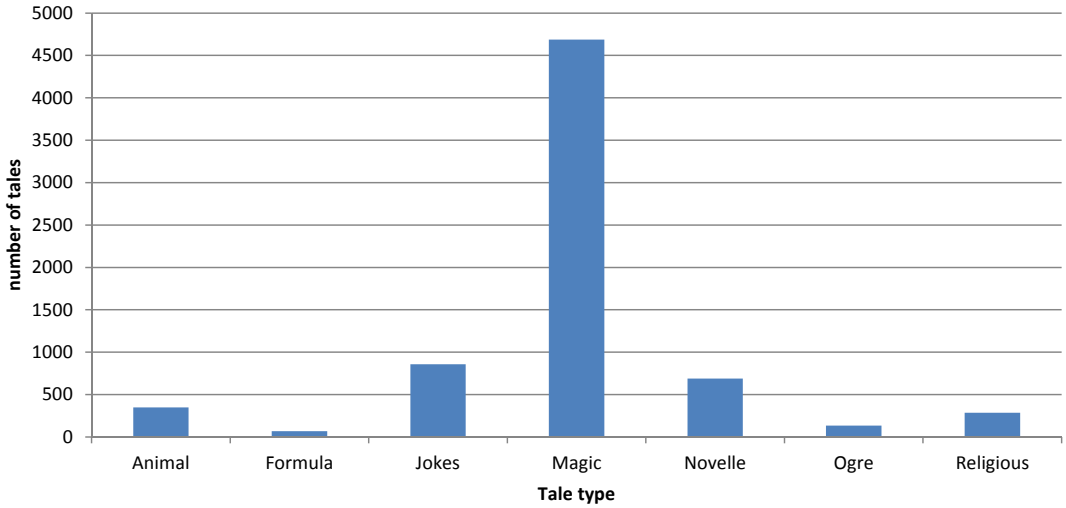

Figure 3 Number of tale types and subtypes per genre.

5.2 Preprocessing

In order to employ sequence mining in a conjunction with machine learning, a preprocessing stage is required. In tale types as motif strings, forkings in the plot may occur as alternative motifs which can be used as filler in particular loci in the plot (see Figure 1). To remove this obstacle, different respective sequences were treated as subtypes (motif string variants) of the same type with a renumbered identifier constructed from the type number and the variant number. This increased the number of 2249 published types to 7050 types and subtypes.

To reflect their similarities, every motif is encoded by a set of unique characters in a certain format in the *Motif Index*. For example, since motifs *"Blindness miraculously cured"* and *"Cripple marvelously cured"* share similar content, they are represented by a similar code having the same prefix letter and a similar number: F952 and F953, respectively. Not to overfit the learning process, we had to generalize motifs in tale types as their sequences and represent them in a less granular way. Decimal motif numbers were gradually truncated, from full notation to class tags only. Due to this process, each level of granularity reduction has fewer number of sequences; however, we do not regard them as new sets of tales, instead, each truncated motif sequence is still considered as another representation of its original sequence. In more details, for every motif, we employed two types of generalization: (1) On a first level we considered only the prefix letter and the integer number to the left of the dot symbol. By doing so, we aggregated similar tales to their father node in the *Motif Index*. In a similar manner, e.g., the first level representation of the two motifs B143.0.3 *"Owl as prophetic bird"* and B143.0.4 *"Raven as prophetic bird"* is generalized into motif B143 *"Prophetic bird"*. (2) Prefix generalization – every motif became represented by its prefix letter only; e.g., the above two motifs both were represented by the letter 'B'. As a result, each original sequence is represented by three sequences: the original, truncation to first level, and truncation to prefix. The principle of this process is displayed in Figure 4.

5.3 Constructing a dictionary of sequences

In machine learning, every instance (tale) is represented by a set of features. We hypothesized that features that are based on sequence frequencies are beneficial for the training process;

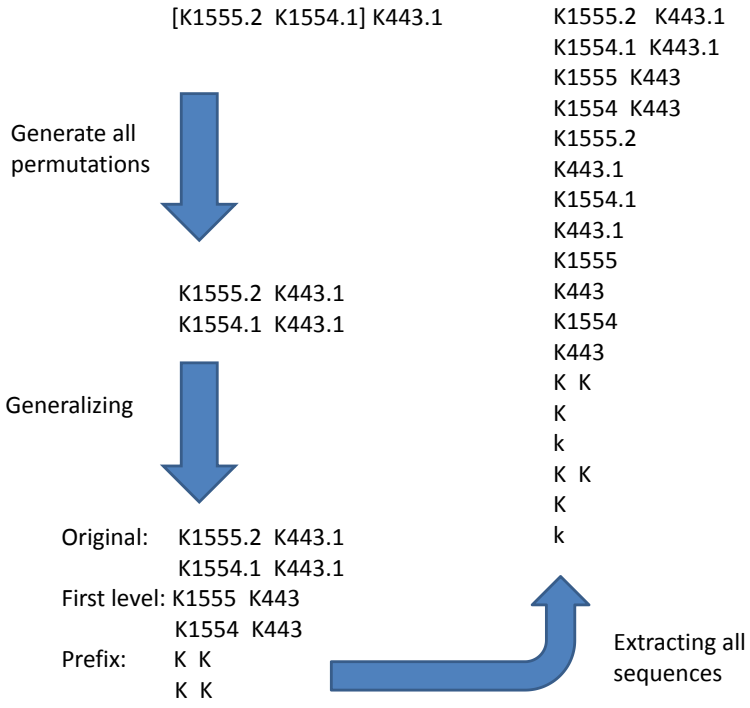

Figure 4 Preprocessing of tale 1358A and extraction of its motif sequences.

therefore, in our work the set of features was extracted from a dictionary of sequence frequencies.

We extracted subsequences from every motif sequence (tale type) and stored their frequencies according to the classes they represented. Since motif sequences are relatively short, we chose to store all sequences of size 1 to 4. Thereby, we avoided storing relatively long sequences that seldom occur and are not likely to add any useful information to our analysis. This is done for each of the three types of sequence representation i.e., the original, and the two generalization rounds. The right side of Figure 4 details subsequence extraction. An example of dictionary entries is given in Table 1.

It is important to note that our dataset contains permutations of motif sequences, as explained in the pre-processing section. As such, they are dependent on their original motif sequence. For example the two permutations in Figure 4 both contain motif K443.1 in their

Table 1 Frequency of motif sequences per tale genre in the dictionary (excerpt).

sequence	tale type	frequency
B184.1 D961 B435.1 H1242	magic	0.0053
L161	magic	0.0498
L161	novelle	0.00182
J H	magic	0.00027
J H	jokes	0.0029
J H	novelle	0.6577

Table 2 Tale types in our corpus, after generating all permutations.

Tale type class	Number of instances
animal	349
formula	68
jokes	858
magic	4668
novelle	688
ogre	133
religious	286

second position. Given that, in our evaluation we ensure that permutations (subtypes) of any original tale type should be assigned to only one set – test or train. If a tale was sampled for the training set, then all of its permutations also belonged to the same set. Only the training set was used to generate the dictionary. Thus, the dictionary does not contain information based on tales from the test set, to avoid dependent tales being used for training.

5.4 Experiment design

After preprocessing and constructing the dictionary, we wish to train a classification model by using sequence discrimination ratios based on statistics.

The original dataset contains seven classes of tale types. Table 1 displays the number of instances for each.

The learning process of any machine learning algorithm requires to be provided with a sufficient number of sampled instances which represent the population of each class. Thus, an effective learning process can be employed. However, in our dataset, in some classes of tale types there are only few dozens of observed instances. Not only having insufficient number of training instances of some classes of tale types, the dataset is also imbalanced.

In a classification problem, class imbalance occurs when there are more examples of a certain class than of any other, on a large scale. For a variety of reasons, imbalanced datasets pose difficulties for induction algorithms [2]. The most obvious problem is that standard machine learning techniques are overwhelmed by the majority class and ignore the minority class. This problem is reflected in the phrase: like a needle in a haystack. Much more than the needle is being small, the problem is the fact that the needle is obscured by a huge number of strands of hay. Therefore, a classifier can achieve high accuracy by always predicting the majority class, particularly if the majority class constitutes most of the dataset, as for the 'magic' class type. Some class labels have a relatively low number of instances, sometimes down to a ratio of 69 less times than in other classes. Therefore, we experimented with those classes of types (i.e., genres) that have a more balanced number of instances.

In our first experiment we tried to discriminate between tales of two type classes, 'jokes' and 'novelle' which have a similar number of instances. The next experiment contains the 'magic' tale types, since this type has the largest number instances that allow an effective learning process. In addition, the mentioned two tale types ('jokes' and 'novelle') were selected as the closest in our dataset to the 'magic' type in terms of number of training instances.

In our classification task, each instance is a sequence of motifs. The goal was to train a model that can be used to predict the class of any unlabeled instance. The first step was the

Table 3 The dictionary as a lookup table of sequences and frequency-ratio, for any given class type.

sequence	tale type	frequency ratio
B184.1 D961 B435.1 H1242	magic	6.89
L161	magic	63.13
L161	novelle	0.041
J H	magic	0.0009
J H	jokes	0.0339
J H	novelle	911.9

extraction of features by which the instances could be represented. To take into consideration the order of motifs, but at the same time also to avoid a rigid structure of motif sequence, we segmented every sequence into several subsequences as detailed in the previous section. We used the subsequences dictionary in the following way. We calculated for every entry its likelihood in each class in contrast with all other classes. This ratio was calculated by dividing the frequency of the subsequences in a specific class by its frequency in all other classes. The calculated frequency ratio was stored in the dictionary, which now functions as a frequency-ratio lookup table (Table 3). We expected that a class whose value was high for a certain subsequence is more likely to be the tale type of a motif sequence that contained this subsequence.

In order to construct the dictionary, in the next step, for every instance we extracted all of its subsequences of size 1 to 4 as a pool of subsequences (see right side of Figure 4). By extracting subsequences of size 1 to 4, we took into consideration the order of the motif sequence, to some extent, while focusing on relatively short subsequences that would more likely to occur in motif sequences, and to avoid overfitting.

We computed two types of features for each class type. First, for all the subsequences from the instance's pool, we attached their frequency-ratio from the lookup table. Then, we sorted them and got the ratio of the highest score. This is the first feature type, which is called 'top 1'. The second is an accumulation of the top three ratios, denoted as 'top 3'. This was repeated for the original motif sequence, and for both generalization rounds of the motif identifiers, i.e., beyond the original, truncation to first level, and truncation to prefix. The total number of features is given by: $2\ (top\ 1\ and\ top\ 3\ ratios) \times 3\ (generalization\ levels) \times class\ types$ (Figure 4). By using abstracted top ratios features, we try to avoid overfitting, as could be the case in a bag-of-motifs feature space approach.

We believe that features that are strongly related to the actual (true) class type of the tale will have higher values than the same features for other class types.

In each experiment we split the dataset into two sets: 80% of the examples were used for training and 20% for testing. The dictionary and its subsequences statistics were also constructed only according to the training instances.

We trained a classification model based on the training set to predict the actual class of each tale. We evaluated several types of machine learning classification algorithms that we find adequate for that task. We chose to display results for the Bayes Network classifier since it yielded the best results, and for a decision tree as a comparison and for illustrating its interpretable model. To train the models and perform the experiments, the WEKA machine learning program suite was used [33].

Table 4 A set of 18 features calculated for each instance.

	Top 1 ratio			Top 3 ratios		
Class / Generalization	Magic	Jokes	Novelle	Magic	Jokes	Novelle
Original	88.5	0	106.60	88.5	0	106.6
First level	118.5	0	136.6	118.5	0	136.6
Prefix	1208	0.11	985	1208	0.11	985

6 Results

The best results are given by Bayes Network classifier. Table 5 details the performance.

In the next experiment we added the 'magic' class instances as well, and the task is to discriminate among the three class types. The results are given by Table 3. The best performance is given by using our approach with the Bayes Network algorithm which outperforms the decision tree algorithm as it yielded better result in more tale types, and across all measurements. We compared our methodology with a baseline. In the baseline, we used a bag-of-words approach, i.e., each tale is represented by a feature vector which is its set of motifs. The best results for the baseline were given after generalizing the motifs to their first-level and by using a decision tree classifier. Our approach outperforms the baseline by all measurements. That is since our approach uses abstracted features, and the baseline uses a set of nearly 2000 features (motifs) that might cause an overfitting. Since the 'magic' class has a substantially higher number of instances, and in order to show performance on each tale type separately, we evaluated each class separately and not the weighted average of the measurements. We analyze the error of the triplet classification experiment. On average, the normalized error rate by motif sequence length is 4.7%, taking into consideration only prominent lengths. Tales of lengths one, four, six and twelve were the most difficult to classify, and resulted in 7%–9% classification errors. Jokes and magic tales were confused

Table 5 Results for binary class experiment. The Bayes Network classifier outperformed the decision tree by F-measure and AUC for both tale types. Best results for each tale type are in bold.

Classifier	Class	Precision	Recall	F-measure	AUC
Bayes Network	novelle	**0.912**	0.601	**0.725**	**0.867**
	jokes	0.749	**0.953**	**0.839**	**0.867**
Decision Tree	novelle	0.463	**0.964**	0.626	0.537
	jokes	**0.783**	0.105	0.185	0.537

Table 6 Results for trinary class experiment. The Bayes Network classifier is found to be superior. Best results for each tale type are in bold.

classifier	class	Precision	Recall	F-measure	AUC
Bayes Network	magic	**0.97**	0.796	0.875	**0.935**
	novelle	0.223	**0.767**	**0.345**	**0.799**
	jokes	**0.844**	**0.409**	**0.551**	0.864
Decision Tree	magic	0.803	**0.983**	**0.884**	0.913
	novelle	**0.6**	0.175	0.271	0.614
	jokes	0.55	0.069	0.123	**0.921**
Decision Tree	magic	0.883	0.056	0.105	0.515
	novelle	0.0	0.0	0.0	0.506
*baseline	jokes	0.136	0.981	0.239	0.518

to be novelles, at almost all error cases; novelles confused to be magic for 70% of its errors. That is for the Bayes Network classifier. However, for the decision tree there is not a same tendency, therefore we can not state that tales are mostly confused to be novelles.

Since the decision tree model is easy to be described and is interpretable, we will explain its structure. The generated tree is a directed graph that consists of a root node (a starting point), internal nodes (nodes that are pointed at and point to other nodes) and leaves (ending points). During the classification process, the classified item "travels" from the root to one of the leaves, where a classification decision is made. Figure 5 illustrates a sub-tree of the generated binary decision tree model. If the 'top 1' subsequence ratio for class 'jokes' in the original pool of subsequences is greater than 4.621245 and the 'top 1' ratio of for class 'magic' in the first level pool of subsequences is not greater than 5.174637 and the 'top 1' ratio for class 'novelle' in the prefix pool of subsequences is not greater than 2.724305, then the instance is of class 'jokes'. The support for this decision is 22/2, based on the training instances.

7 Conclusion and future research

Considering the existence of "narrative DNA", we used sequence mining methods used in bioinformatics to learn more about the nature of tale types as a corpus. 94% of the Aarne-Thompson-Uther catalogue (2249 tale types in 7050 variants) was analyzed as individual motif strings based on the *Motif Index* and scanned for similar subsequences. Next, using a machine learning classification algorithm, we built and evaluated a classifier which predicts the tale type of a new motif sequence. Our findings indicate that the probabilistic underpinnings of tale types by motif co-occurrences are robust enough to develop the classification model which, on this instance, was able to predict motif strings characterizing magic tales, novelles and jokes. We plan to continue this work and combine our framework with sequence transformation analysis to learn more about the DNA-like nature of narrative content.

Acknowledgements. The authors are grateful to two unknown reviewers for their observations and suggestions.

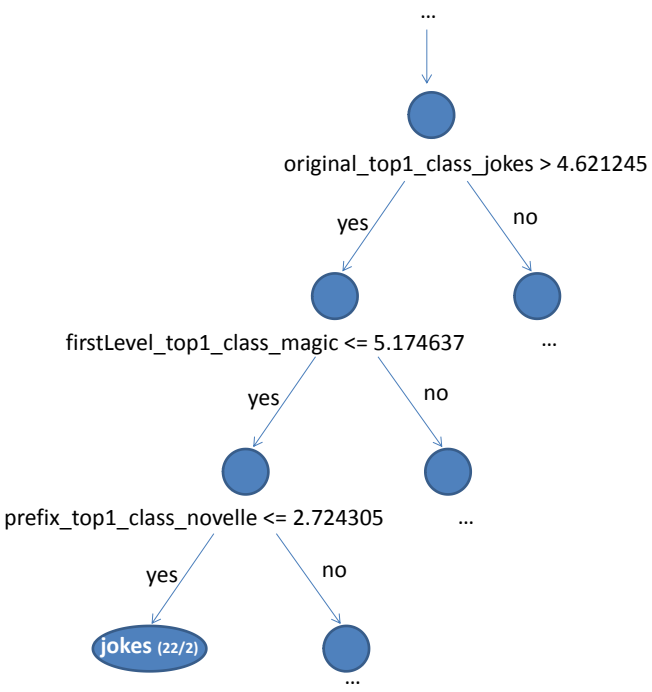

Figure 5 A sub-tree of a decision tree model, from the root note to one of its decision leaves.

References

1 M. Berti. The Ancient Greek and Latin dependency treebanks. blog post, `http://www.monicaberti.it/2010/10/the-ancient-greek-and-latin-dependency-treebanks/`, 2010.

2 N. V. Chawla, N. Japkowicz, and A. Kołcz. Editorial: special issue on learning from imbalanced data sets. *ACM SIGKDD Explorations Newsletter*, 6(1):1–6, 2004.

3 C. Danescu-Niculescu-Mizil, J. Cheng, J. Kleinberg, and L. Lee. You had me at hello: How phrasing affects memorability. Preprint, 2012.

4 S. Darányi. Examples of formulaity in narratives and scientific communication. manuscript, University of Szeged, Hungary, 2010.

5 S. Darányi and L. Forró. Detecting multiple motif co-occurrences in the Aarne-Thompson-Uther tale type catalog: A preliminary survey. *Anales de Documentación*, 15(1), 2012.

6 S. Darányi, P. Wittek, and L. Forro. Toward sequencing "narrative DNA": Tale types, motif strings and memetic pathways. In Mark Alan Finlayson, editor, *Proceedings of Computational Models of Narrative 2012*, pages 2–10, İstanbul, 2012.

7 R. Dawkins. *The Selfish Gene*. Oxford University Press, Oxford, 1976.

8 G. Dong. *Sequence data mining*. Number 33 in Advances in Database Systems. Springer, 2009.

9 A. Dundes. The motif-index and the tale type index: A critique. *Journal of Folklore Research*, 34(3):195–202, 1997.

10 P. Ferreira and P. Azevedo. Protein sequence classification through relevant sequence mining and Bayes classifiers. In Carlos Bento, Amílcar Cardoso, and Gaël Dias, editors, *Progress in Artificial Intelligence, 12th Portuguese Conference on Artificial Intelligence,*

EPIA 2005, Covilhã, Portugal, December 5–8, 2005, Proceedings, number 3808 in Lecture Notes in Computer Science, pages 236–247. Springer, 2005.

11 A. Garnham. What is wrong with story grammars. *Cognition*, 15:145–154, 1983.

12 M. Griffin. An expanded, narrative algebra for mythic spacetime. *Journal of Literary Semantics*, 30:71–82, 2001.

13 M. Griffin. More features of the mythic spacetime algebra. *Journal of Literary Semantics*, 32:49–72, 2003.

14 M. Griffin. Mythic algebra uses: Metaphor, logic, and the semiotic sign. *Semiotica*, 158–1/4:309–318, 2006.

15 Z. S. Harris, M. Gottfried, T. Ryckman, P. Mattick, A. Daladier, T. N. Harris, and S. Harris. *The form of information in science: analysis of an immunology sublanguage.* Kluwer, Dordrecht, 1989.

16 H. Jason. *Motif, Type and Genre. A Manual for Compilation of Indices and A Bibliography of Indices and Indexing.* Academia Scientiarum Fennica, Helsinki, 2000.

17 H. Jason and D. Segal, editors. *Patterns in oral literature.* Mouton, The Hague, 1977.

18 M. L. Jockers. *Macroanalysis: Digital methods and literary history.* University of Illinois Press, Champaign, Il., 2013.

19 Folgert Karsdorp, Peter Van Kranenburg, Theo Meder, Dolf Trieschnigg, and Antal Van den Bosch. In search of an appropriate abstraction level for motif annotations. In Mark Alan Finlayson, editor, *Proceedings of Computational Models of Narrative 2012*, pages 22–26, İstanbul, 2012.

20 G. P. Lakoff. Structural complexity in fairy tales. *The Study of Man*, I:128–190, 1972.

21 C. Lévi-Strauss. *Mythologiques I–IV.* Plon, Paris, France, 1964–1971.

22 A. Lord. *The singer of tales.* Harvard University Press, Cambridge, 1960.

23 P. Maranda. *The double twist: from ethnography to morphodynamics.* University of Toronto Press, Toronto, 2001.

24 M. Parry. Studies in the epic technique of oral verse-making. I: Homer and Homeric style. *Harvard Studies in Classical Philology*, 41:73–143, 1930.

25 M. Parry. Studies in the epic technique of oral verse-making. II: The Homeric language as the language of an oral poetry. *Harvard Studies in Classical Philology*, 43:1–50, 1930.

26 Vladimir Yakovlevich Propp. *Morphology of the Folktale.* University of Texas Press, Austin, TX, 2nd edition, 1968. transl. L. Scott.

27 S. R Ranganathan. *The Colon Classification Vol. 4.* Rutgers Graduate School of Library Service, New Brunswick, NJ, 1965.

28 N. Roshianu. *Traditsionnuie formuly skazki (Traditional formulae of the fairy tale).* Moscow, 1974.

29 R. M. Ross, S. J. Greenhill, and Q. D. Atkinson. Population structure and cultural geography of a folktale in Europe. *Proceedings of the Royal Society B*, 280(1756):1471–2954, 2013.

30 S. Thompson. *Motif-index of Folk-Literature: a Classification of Narrative Elements in Folktales, Ballads, Myths, Fables, Medieval Romances, Exempla, Fabliaux, Jest-Books, and Local Legends. 6 volumes.* Indiana University Press, Bloomington, 2nd edition, 1955–1958.

31 J. Trier. Das sprachliche Feld. *Neue Jahrbücher für Wissenschaft und Jugendbildung*, 10:428–449, 1934.

32 H. J. Uther. *The types of international folktales: A classification and bibliography based on the system of Antti Aarne and Stith Thompson.* Academia Scientiarum Fennica, Helsinki, Finland, 2004.

33 I. H. Witten and E. Frank. *Data Mining: Practical machine learning tools and techniques.* Morgan Kaufmann, 2005.

A Appendix: The Types of International Folktales (from [32])

1. ANIMAL TALES (1–299)
Wild Animals 1–99
    The Clever Fox (Other Animal) 1–69
    Other Wild Animals 70–99
Wild Animals and Domestic Animals 100–149
Wild Animals and Humans 150–199
Domestic Animals 200–219
Other Animals and Objects 220–299
    Birds 220–249
    Fish 250–253

2. TALES OF MAGIC (300–749)
Supernatural Adversaries 300–399
Supernatural or Enchanted Wife (Husband) or Other Relative 400–459
    Wife 400–424
    Husband 425–449
    Brother or Sister 450–459
Supernatural Tasks 460–499
Supernatural Helpers 500–559
Magic Objects 560–649
Supernatural Power or Knowledge 650–699
Other Tales of the Supernatural 700–749

3. RELIGIOUS TALES (750–849)
God Rewards and Punishes 750–779
The Truth Comes to Light 780–799
Heaven 800–809
The Devil 810–826
Other Religious Tales 827–849

4. REALISTIC TALES (NOVELLE) (850–999)
The Man Marries the Princess 850–869
The Woman Marries the Prince 870–879
Proofs of Fidelity and Innocence 880–899
The Obstinate Wife Learns to Obey 900–909
Good Precepts 910–919
Clever Acts and Words 920–929
Tales of Fate 930–949
Robbers and Murderers 950–969
Other Realistic Tales 970–999

5. TALES OF THE STUPID OGRE (GIANT, DEVIL) (1000–1199)
Labor Contract 1000–1029
Partnership between Man and Ogre 1030–1059
Contest between Man and Ogre 1060–1114
Man Kills (Injures) Ogre 1115–1144
Ogre Frightened by Man 1145–1154
Man Outwits the Devil 1155–1169
Souls Saved from the Devil 1170–1199

6. ANECDOTES AND JOKES (1200–1999)
Stories about a Fool 1200–1349

Stories about Married Couples 1350–1439
The Foolish Wife and her Husband 1380–1404
The Foolish Husband and his Wife 1405–1429
The Foolish Couple 1430–1439
Stories about a Woman 1440–1524
Looking for a Wife 1450–1474
Jokes about Old Maids 1475–1499
Other Stories about Women 1500–1524
Stories about a Man 1525–1724
The Clever Man 1525–1639
Lucky Accidents 1640–1674
The Stupid Man 1675–1724
Jokes about Clergymen and Religious Figures 1725–1849
The Clergyman Is Tricked 1725–1774
Clergyman and Sexton 1775–1799
Other Jokes about Religious Figures 1800–1849
Anecdotes about Other Groups of People 1850–1874
Tall Tales 1875–1999

7. FORMULA TALES (2000–2399)
Cumulative Tales 2000–2100
Catch Tales 2200–2299
Other Formula Tales 2300–2399

B Appendix: Excerpt from Thompson's Motif Index [30]

B0–B99.Mythical animals
B10. Mythical beasts and hybrids
B20. Beast-men
B30. Mythical birds
B40. Bird-beasts
B50. Bird-men
B60. Mythical fish
B70. Fish-beasts
B80. Fish-men
B90. Other mythical animals
B100–B199. Magic animals
B100–B119. Treasure animals
B100. Treasure animals–general
B110. Treasure-producing parts of animals
B120–B169. Animals with magic wisdom
B120. Wise animals
B130. Truth-telling animals
B140. Prophetic animals
B150. Oracular animals
B160. Wisdom-giving animals
B170–B189. Other magic animals
B170. Magic birds, fish, reptiles, etc.
B180. Magic quadrupeds
B190. Magic animals: miscellaneous motifs
B200–B299. Animals with human traits
B210. Speaking animals
B220. Animal kingdom (community)

B230. Parliament of animals
B240. King of animals
B250. Religious animals
B260. Animal warfare
B270. Animals in legal relations
B280. Animal weddings
B290. Other animals with human traits
B300–B599. Friendly animals
B300–B349. Helpful animals–general
B310. Acquisition of helpful animal
B320. Reward of helpful animal
B330. Death of helpful animal
B340. Treatment of helpful animal—miscellaneous
B350–B399. Grateful animals
B360. Animals grateful for rescue from peril of death
B370. Animals grateful to captor for release
B380. Animals grateful for relief from pain
B390. Animals grateful for other kind acts
B400–B499. Kinds of helpful animals
B400–B449. Helpful beasts
B400. Helpful domestic beasts
B430. Helpful wild beasts
B450. Helpful birds
B470. Helpful fish
B480. Helpful insects
B490. Other helpful animals
B500–B599. Services of helpful animals
B500. Magic power from animal
B510. Healing by animal
B520. Animals save person's life
B530. Animals nourish men
B540. Animal rescuer or retriever
B550. Animals carry men
B560. Animals advise men
B570. Animals serve men
B580. Animals help men to wealth and greatness
B590. Miscellaneous services of helpful animals
B600–B699. Marriage of person to animal
B610. Animal paramour
B620. Animal suitor
B630. Offspring of marriage to animal
B640. Marriage to person in animal form
B650. Marriage to animal in human form
B700–B799. Fanciful traits of animals
B710. Fanciful origin of animals
B720–B749. Fanciful physical qualities of animals
B720. Fanciful bodily members of animals
B730. Fanciful color, smell, etc. of animals
B740. Fanciful marvelous strength of animals
B750. Fanciful habits of animals
B770. Other fanciful traits of animals
B800–B899. Miscellaneous animal motifs
B870. Giant animals